Transforming Voice Quality

Transforming Voice Quality

Ben Gillett, now with Camel Audio (www.camelaudio.com)
Simon King, Centre for Speech Technology Research, University of Edinburgh, UK

ben@camelaudio.com, Simon.King@ed.ac.uk

Introduction

Goal

Voice transformation is the process of transforming the characterstics of a source speaker, such that a listener would believe the speech was uttered by some target speaker

- Need to transform both
  1. voice quality
     - source characteristics
     - vocal tract frequency response
  2. intonation
     - F0
     - segment durations
     - amplitude

This poster covers the first topic. In session PMoCe we presented a simple F0 transformation method.

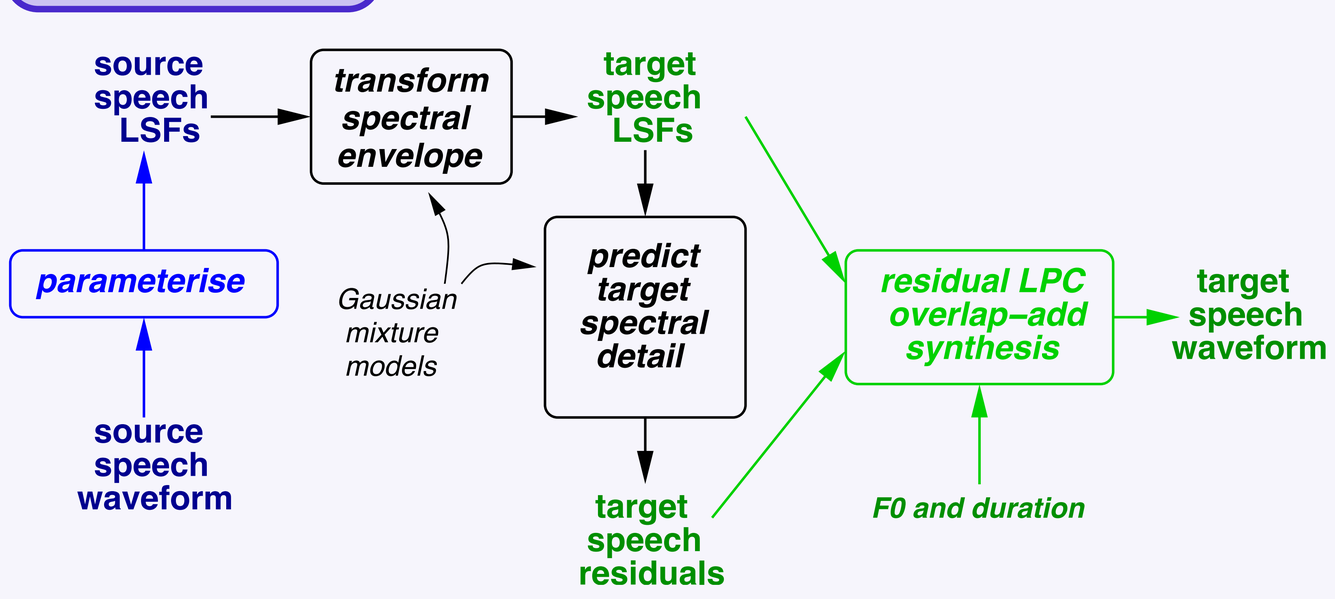
Why?

Applications include:

- speech synthesis
  - unit selection voices are expensive to construct: if we could transform existing voices easily, then we could make many new voices quickly and cheaply
  - could construct voices based on speakers for which we have only small speech samples
  - could modify existing voices in subtle ways (more breathy, richer, smoother, more sexy, ...)
- low bit-rate speech coding
  - Tx sends speaker information once, then a stream of segmental information; Rx resynthesises
- entertainment
- voice disguise

Overview

System diagram



Data required

To train the two mappings (which will be Gaussian mixture models), we require aligned speech from source and target speakers

- source and target speaker reading the same text
- dynamic-time-warping to align frame-by-frame
- data pruning to remove badly aligned frames (see below)

Spectral envelope transform

Novel part: data pruning

Why prune data?

- training needs aligned pairs of frames of source and target speech
- badly aligned frames introduce noise and lead to poor models

Pre-GMM pruning

Remove all pairs of frames that

- do not match in voicing (one voiced, other unvoiced)
- have large amplitude mismatches (details in Gillett's thesis)

together, this removes about 25% of frame pairs.

Post-GMM pruning

Remove all pairs of frames that are the least probable under the GMM. This removes 15% of the remaining frames. GMM is retrained on remaining frames. 15% was determined empirically through informal listening tests.

Method

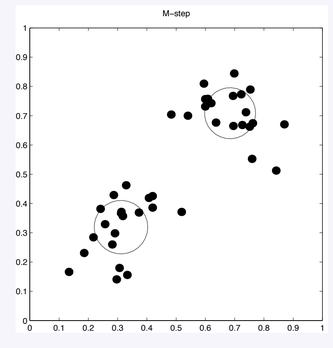
After Kain

Method is basically the same as that of Kain, but

- we prune the training data (see left)
- we remove the restriction that the speech is spoken in a monotone
- this introduces a new problem into spectral detail transformation – variable frame lengths – see above right

GMM

- Train: aligned pairs of frame source and target speech, parameterised as **Line spectral frequencies** – good interpolation properties.
- Predict: give only source LSFs, predict target LSFs



Spectral detail prediction

Overview

Only for voiced frames

- voiced frames: speaker identity in residual, so predict it
- unvoiced frames: little speaker identity, so use source residual

Problem: variable frame lengths

Pitch-synchronous analysis means variable frame lengths
Hence, residuals (time or frequency domain) for each frame vary in length
Using a Gaussian **mixture** model requires computing weighted sums
Which requires frame length normalisation, i.e. re-sampling

Magnitude spectrum

Gaussian mixture model predicts y given x thus

$$E[y|x] = \sum_{q=1}^{Q}(\mu_q^Y + \Sigma_q^{YX}(\Sigma_q^{XX})^{-1}(x - \mu_q^X)).P(c_q|x)$$

where

$$\Sigma_q = \begin{bmatrix} \Sigma_q^{XX} \Sigma_q^{XY} \\ \Sigma_q^{YX} \Sigma_q^{YY} \end{bmatrix}$$

$$\mu_q = \begin{bmatrix} \mu_q^X \\ \mu_q^Y \end{bmatrix}$$

$$P(c_p|x) = \frac{\alpha_p N(x; \mu_p; \Sigma_p)}{\sum_{q=1}^{Q} \alpha_q N(x; \mu_q; \Sigma_q)}$$

and α_q are the mixture weights of the Q multivariate Gaussians in the GMM.

In the case of residual magnitude spectrum prediction, x are the previously predicted target LSFs and y is the magnitude spectrum being predicted.

This requires y to have a constant size for all frames – that's why resampling is required.

(In practice, x is actually a vector of cepstral coefficients derived from the LSFs)

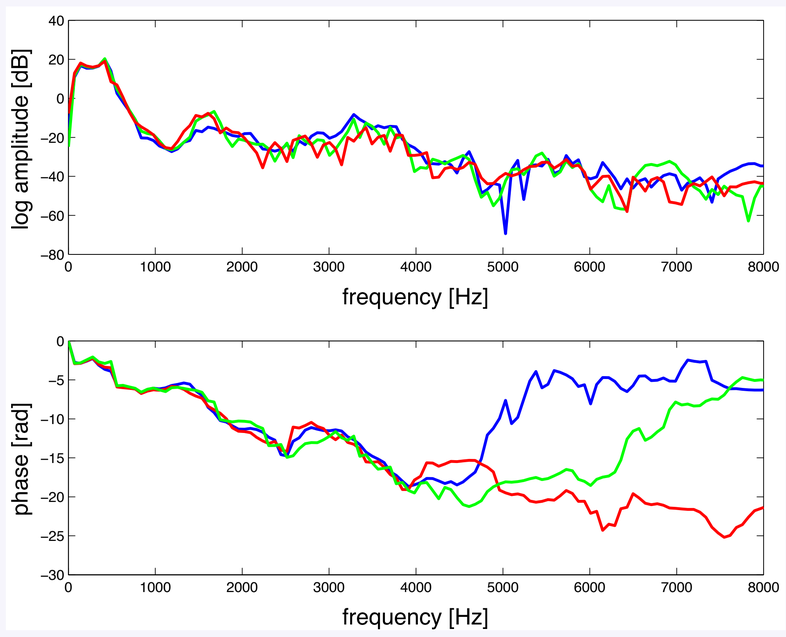
Phase spectrum

Phase spectrum is not as suitable for re-sampling

- phase is highly sensitive to alignment between analysis window and moment of glottal closure
- phase is wrapped: unwrapping is error-prone
- phase spectrum at higher frequencies (above 4kHz) is inconsistent from frame to frame.

The combination of the above two problems means that length-normalisation (i.e. re-sampling) of residual phase spectrum is not easy.

Figure shows three consecutive pitch-synchronous frames of voiced speech - note how phase is slowly varying frame-to-frame only up to 4kHz. Phase slope has been corrected for alignment variation.



This means we cannot use a GMM in the usual way to predict the phase spectrum

Instead, we must use an actual residual. In the GMM, we keep the target residual phase spectra for a selection of the most probable data points. At synthesis time, we pick an actual phase spectrum that is as close to the required length as possible. No resampling is performed on the phase spectrum.

Evaluation

Performance measures

Kain's performance index can be used to evaluate the spectral envelope transform $P_{LSF} = 1 - \frac{E_{LSF}(t(n),\hat{t}(n))}{E_{LSF}(t(n),s(n))}$
Kain's system has $P_{LSF} = 0.31$ and our system has $P_{LSF} = 0.36$ although on different data (we believe our data set to be more challenging as it is prosodically varied)

Our data: 2 male and 2 female speakers of the Boston University Radio Corpus. 2 minutes of training data and 1 minute of test data per speaker. Transforms were only between same sex pairs; figure above is average over transforms: f1a $\rightarrow$ f2b, f2b $\rightarrow$ f1a, m1a $\rightarrow$ m2b and m2b $\rightarrow$ m1a.

Audio examples

Summary

What next?

- reducing the signal processing artefacts
  - output currently has too many artefacts – typical of RELP and LP-PSOLA synthesis
  - we think many of these problems are fixable
- trying the system with a *lot* more data, e.g. a pair of unit-selection synthesis voices (hours of speech, not minutes)
- morphing, rather than complete transformation
  - smoothly varying the output between a pair of speakers
  - interpolating between more than two speakers
  - creating "new" speakers from existing ones
- including phonetic transcription into process (e.g. to get better alignment of source and target training data) – transcriptions are available for unit-selection voices

See also...

- at this conference:
  - poster by Gillett & King in session PMoCe (a simple F0 transformation method)
  - posters by Shiga & King in sessions PMoCg and PWeBe (improved source and filter estimation using multiple frames)
- **`www.cstr.ed.ac.uk`** for latest progress on voice transformation and speech synthesis
- **`www.camelaudio.com`** for musical instrument transformation and morphing